7-1-1992

Full Text on CD-ROM

Carol Tenopir

University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Full Text on CD-ROM

[This column is based on a speech presented at the 1992 Annual Meeting of the Special Libraries Association in San Francisco.]

AS THE USE of bibliographic tools increases in libraries thanks to popular electronic versions of old indexing/abstracting standbys, libraries need more full texts to support increased demand. CD-ROM is a natural medium to supplement or replace paper collections or full texts online.

The issues and trends surrounding full text on CD-ROM can be categorized by the journalist's dictum of the five Ws and an H (slightly scrambled for our purposes).

How many full texts are there?

What types are there?

Who is producing full texts?

Why is careful decision-making for purchasing so important?

Where are full texts heading?

When will full text on CD-ROM surpass online or print, and when will other changes occur?

How many full texts are there?

First, *how* many full texts are there to choose from in 1992? It won't surprise you that there are more than there were in 1991 and a lot more than in 1990. Electronic full texts are growing at a faster pace than any other type of database in both the online and CD-ROM media.

The premiere issue (January 1990) of the *Directory of Portable Databases* listed a total of 409 CD-ROM databases. About one-quarter of these (100) were full text in whole or in part. (Often databases are categorized as more than one type: full text/bibliographic, for example. I counted mixed type as full text.) By the October 1991 issue, the total number of CD-ROM databases had jumped to 934—319 (34 percent) of which were full text. The latest issue (April 1992) lists 1030 CD-ROM databases—382 (37 percent) of which are full text.

The important message in these figures is not just that CD-ROM full texts are enjoying rapid growth, but also that there is a greater variety of titles from which to choose. It means if something you want is not available now, it might be in six months or next year or the year after. The important thing about *how* is, therefore, *how* to find out on a regular basis what new titles and types of full text are available.

What types of full text are on CD?

That brings me to my next W—*what*. *What* types of full texts can you purchase on CD-ROM? There are two dimensions to the *what* question: 1) *What* subjects are available and 2) *What* kind of text do they include?

Full texts on CD-ROM cover almost every type of text, just as print and online texts do, but certain types lend themselves better to CD-ROM. Very timely materials such as newswires, for example, are found widely online, but they are less common on CD where their currency is lost. Texts that rely heavily on graphics, on the other hand, are not yet suitable for online but work well for CD-ROM (or print). Texts that benefit from a high degree of interactivity are naturally suited to CD-ROM.

The other dimension of what

The other dimension of *what* is *what* do these texts include or what searching powers do they support? The jury is still out with regard to CD-ROM full texts, but right now there are two main choices: 1) ASCII (text) files, and 2) Image files.

ASCII files—just like online databases—provide only the textual portions of documents, without graphics or fancy typefaces, but every word of the text can be searched, and the text can be downloaded and manipulated. ASCII texts are usually rekeyed by the database producer or scanned and converted into machine-readable text with optical character recognition (OCR) software. ASCII versions of magazines or newspapers often include only selected articles from each source, and rarely include nonarticle portions such as letters to the editor, advertisements, announcements of meetings, and so on.

Image files are scanned, but the texts are not converted to searchable text with OCR software. Instead, both the graphics and text are treated as images. (Letters are just a pattern of light and dark that forms a shape, just like photographs or other images.) Image files give you exact images of the original document (much as micro-

Carol Tenopir is Associate Professor at the School of Library and Information Studies, University of Hawaii at Manoa, Honolulu

Full-Text Databases

The six types of full texts that have the most CD-ROM titles are, in rank order by number of products:

1. Books—many reference books such as encyclopedias, dictionaries, thesauri, etc., but also fiction such as Sherlock Holmes, American authors, novels, children's books, and more. Many are multimedia.

2. Government documents and reports—e.g., patents, SEC filings, and regulations, government publications such as the Code of Federal Regulations and Commerce Business Daily.

3. Laws and legislation—statutes, court decisions, and other primary legal documents from all over the world.

4. Journals and magazines—ranging from general interest to scholarly/technical, from Magazine Rack to IEEE Journals.

5. News—including newspapers and some newswires. Both DIALOG and UMI/ProQuest are providing many daily papers.

6. Unpublished information—e.g., software documentation, unpublished reports, and press releases.

forms do) so you get graphics, layout, page breaks, etc., like the original. You sacrifice the ability to search and manipulate the text.

Actually, a third option is becoming popular: combination products that combine searchable ASCII text *and* scanned graphic images. Although this option takes more space on the CD-ROM, better compression techniques are allowing more to be put onto each disc. Combination products provide the best of both alternatives—the power of searching ASCII text with the aesthetics and completeness of page images.

Why is the decision important?

This leads directly to *why* it is important to make good decisions. You must make decisions because there are now competing products that may not be giving you exactly the same thing. EBSCO's Magazine Article Summaries (MAS) includes ASCII texts of selected magazines, many of which are in image form in UMI's General Periodicals Ondisc. Information Access Company's (IAC) General Periodicals Index includes ASCII journals; UMI's Business Periodicals Ondisc is image.

Why also pertains to the cost of CD-ROM. The good news is that costs are coming down, but, in most cases, CD versions are still more expensive than print, and, unlike online, you must pay up-front whether you subsequently use the title or not.

Paul T. Nicholls's *CD-ROM Collection Builder's Toolkit* (Online, 1991) traces the descending median costs for CD-ROM over the past five years. Median cost for all CDs in 1991 was $702, down from $1,273 in 1987. The median cost of full-text CDs was $630 in 1991. There is much variation in the prices of full texts; many are now well under $100, especially those aimed at the consumer market.

Who produces and sells CD-ROM?

Who is producing these full texts on CD-ROM, whether ASCII or image? Many of the vendors and database producers from online or print are using CD-ROM as another distribution medium, but there are new players as well. More than with any other type of CD-ROM, full-text publishers are likely to be directly involved with both producing and selling their products without an intermediary. They may or may not use their own proprietary software.

There are five main models for CD-ROM creation/production/distribution. In the first model, the producers do everything, being responsible for the work itself, using proprietary software created inhouse, and marketing or distributing their products. (Two examples are Britannica Software, which produces Compton's MultiMedia Encyclopedia, and West Publishing Company's legal products.) The software is often geared to the specific text, which is good when you are searching that text but means new techniques have to be learned for each new title.

The second model is similar to the first except here the producer is not a primary publisher. The database producers obtain the rights to other publishers' works, convert the texts to their software, and often add value such as subject headings, a new arrangement, combinations of titles, and so on. The database producers do all of the marketing and distribution. (Database producers include UMI/ProQuest, Newsbank, IAC/InfoTrac, and CMC ReSearch, Inc.) Each of these producers is likely to have several different products using the same software, and they may provide access to a multitude of texts.

The third model is a variation on the second. In this model, the publisher is the intellectual creator of the material and does the marketing and distribution. The CD-ROM software, however, is leased or modified under contract with a separate software company. This is a common model, used by Bowker for all of its CD products (Online Computer Systems, Inc. developed the software), by Grolier (Grolier now uses Online Computer Systems software for the new editions of its Grolier Electronic Encyclopedia), and by Pergamon (it uses KnowledgeSet software).

Fourth is the vendor/service bureau—companies like DIALOG and SilverPlatter that lease databases, convert them to their software, and sell directly to users. Many titles are necessary for the vendor to make this economical, so the advantage to the user is a variety of titles that use the same software.

The fifth and last model is the distributor or jobber that sells other companies' CD-ROM products, but does no creation, production, or conversion. In fact, the distributor has nothing to do with any aspect of the creation of the product. Updata, Bureau of Electronic Publishing, and Faxon all act as jobbers for other companies' titles. The advantages are the same as those with any book jobber—you deal with just one company for your purchases; the disadvantage is that this one company is only an intermediary that holds no responsibility for the content or the software.

Where and when?

That brings us to our last Ws: *where* and *when*. *Where* is full text on CD-ROM going, and *when* will it get there? Based on projections from the Optical Publishers Association and trends over time, Nicholls predicted that the number of CD-ROM titles would surpass the number of online titles by the end of 1992, and there is no reason to believe it won't happen sooner. The trends point undeniably to many more full texts online and on CD-ROM in the coming years. In addition to standard CD-ROM products, CD full texts are getting diversified with CD-I (compact disc interactive), CDTV, and the Sony Data Discman™—all of which require separate, incompatible hardware.

Besides the increase in the sheer number of products, they will undoubtedly use more graphics, especially with full text. These will be both image files and combination ASCII/Image files. While not every product benefits from motion or sound, multimedia of at least text and graphics is quickly becoming the standard rather than the exception for CD-ROM full texts. These texts will be available over local area networks (LANs) and over dial-up wide area networks (WANs), both from within your organization and in cooperative arrangements with other libraries.

It is safe to say *where* we are going with full text on CD-ROM is toward more products, more text, more graphics, more types of texts. It is also safe to say the *when* is now.

In libraries, full text on CD-ROM can one day replace a large part of the current print collection, plus provide access to resources for which you currently don't have space. What it all means is more choices, more products and software to evaluate, and, as always, more ways to spend your money.

[Ed. note: *LJ*'s CD-ROM Reviews column debuts in this issue on p. 136. Next month, CD-ROM Reviews will include some full-text products.]